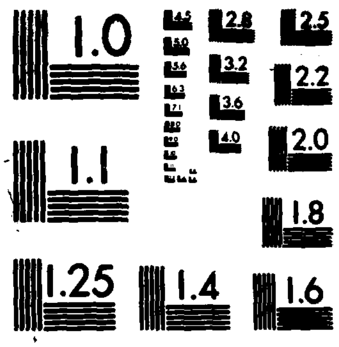
AD-A171 938   DIAGNOSTICS AND ROBUST ESTIMATION WHEN TRANSFORMING THE   1/1
REGRESSION MODEL A.. (U) NORTH CAROLINA UNIV AT CHAPEL
HILL DEPT OF STATISTICS R J CARROLL ET AL. OCT 85

UNCLASSIFIED MMS-1592 AFOSR-TR-86-0682 F49620-85-C-0144 F/G 12/1   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

8c. ADDRESS (City, State and ZIP Code)

Bolling Air Force Base
Washington, DC 20332

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2403	A5	

11. TITLE (Include Security Classification)

"Diagnostics and Robust Estimation When Transforming the Regression Model and the Responses"

12. PERSONAL AUTHOR(S)

Carroll, R.J. and Ruppert, David

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
technical	FROM 8/85 TO 8/86	October 1985	38

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

parametric transformation, maximum likelihood

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In regression analysis, the response is often transformed to remove heteroscedasticity and/or skewness. When a model already exists for the untransformed response, then it can be preserved by transforming both the model and the response with the same transformation. This methodology, ~~which we call~~ called "transform both sides" has been applied in several recent papers, and appears highly useful in practice. When a parametric transformation family such as power transformations is used, then the transformation can be estimated by maximum likelihood. The MLE however is very sensitive to outliers. ~~In~~ this article, we proposes diagnostics which indicate cases influential for the transformation regression parameters. We also propose a robust bounded-influence estimator similar to the Krasker-Welsch regression estimate. Both diagnostics and the robust estimator can be implemented on standard software.

AFOSR-TR- 86-0682

"DIAGNOSTICS AND ROBUST ESTIMATION WHEN TRANSFORMING THE REGRESSION MODEL AND THE RESPONSES"

by

R.J. Carroll and David Ruppert

Mimeo Series #1592

October 1985

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Diagnostics and Robust Estimation When Transforming The Regression Model and The Response

R. J. Carroll

and

David Ruppert

Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27514

Abstract: In regression analysis, the response is often transformed to remove heteroscedasticity and/or skewness. When a model already exists for the untransformed response, then it can be preserved by transforming both the model and the response with the same transformation. This methodology, which we call "transform both sides" has been applied in several recent papers, and appears highly useful in practice. When a parametric transformation family such as power transformations is used, then the transformation can be estimated by maximum likelihood. The MLE however is very sensitive to outliers. In this article, we propose

diagnostics which indicate cases influential for the transformation or regression parameters. We also propose a robust bounded-influence estimator similar to the Krasker-Welsch regression estimate. Both the diagnostics and the robust estimator can be implemented on standard software.

ACKNOWLEDGEMENT

The research of Professor Carroll was supported by Air Force Office of Scientific Research Contract AFOSR-S-49620-85-C-0144. Professor Ruppert's research was supported by the National Science Foundation Grant DMS-8400602.

1. INTRODUCTION

In regression analysis, the response y is often transformed for two distinct purposes, to induce normally distributed, homoscedastic errors and to improve the fit to some simple model involving explanatory variables $\underline{x}$. In many situations, however, y is already believed to fit a known model $f(\underline{x};\underline{\beta})$ $\underline{\beta}$ being a p-dimensional parameter vector. If a transformation of y is still needed to remove skewness and/or heteroscedasticity, then one can transform both y and $f(\underline{x};\underline{\beta})$ in the same manner. Specifically, let $y^{(\lambda)}$ be a transformation indexed by the parameter λ and assume that for some value of λ

$$Y_i^{(\lambda)} = f^{(\lambda)}(\underline{x};\underline{\beta}) + \sigma\epsilon_i \tag{1}$$

where $\epsilon_1,\ldots,\epsilon_N$ are independent and at least approximately normally distributed. Notice the difference between (1) and the usual approach of transforming only the response, not $f(\underline{x};\underline{\beta})$, i.e.,

$$y^{(\lambda)} = f(\underline{x};\underline{\beta}) + \sigma\epsilon_i. \tag{2}$$

It should be emphasized that model (1) is not a substitute for (2). Both models are appropriate, but under different circumstances. Model (2) has been amply discussed by Box and Cox (1964) and others, e.g., Draper and Smith (1980) and Cook and Weisberg (1982). Typically in model (2), $f(\underline{x};\underline{\beta})$ is linear but in principle nonlinear models can be used. Model (1), which we call "transform both sides", has been discussed extensively in Carroll & Ruppert (1984), and Snee (1985), and Ruppert and Carroll (1986) and we will only summarize those discussions. According to (1), $f(\underline{x};\underline{\beta})$ has two closely related interpretations; $f(\underline{x};\underline{\beta})$ is the value of y when the error is zero and it is the median of the conditional distribution of y given $\underline{x}$. In Carroll and Ruppert (1984), we were concerned with situations where a physical or biological model provides

$f(\underline{x};\underline{\beta})$, but where the error structure is a priori unknown. Examples by Snee (1985), Carroll and Ruppert (1984), Ruppert and Carroll (1986), and Bates, Wolf, and Watts (1985) show that transforming both sides can be highly effective with real data, both when a theoretical model is available and, as Snee shows, when $f(\underline{x};\underline{\beta})$ is obtained empirically.

By estimating λ, σ and $\underline{\beta}$ simultaneously, rather than simply fitting the original response y to $f(\underline{x};\underline{\beta})$, we achieve two purposes. Firstly, $\underline{\beta}$ is estimated efficiently and therefore we obtain an efficient estimate of the conditional median of y. Secondly, we model the entire conditional distribution of y given $\underline{x}$, and, in particular, we have a model wich can account for the skewness and heteroscedasticity in the data. Carroll and Ruppert (1984) discuss the importance of modeling the conditional distribution of y in a special case, a spawner-recruit analysis of the Atlantic menhaden population. To specify the conditional distribution of y, for fixed λ let $h(y,\lambda)$ be the inverse of $y^{(\lambda)}$, i.e. $h(y^{(\lambda)},\lambda) \equiv y$, and let F be the distribution function of ϵ_i. We assume that F is approximately normal, but not necessarily exactly normal since if $y^{(\lambda)}$ is the Box-Cox (1964) modified power family,

$$
\begin{aligned}
y^{(\lambda)} &= (y^{\lambda}-1)/x \qquad && \lambda\neq 0 \\
&= \log(y) && \lambda=0,
\end{aligned}
\tag{3}
$$

then F must have finite support whenever $\lambda\neq 0$. The p-th quantile of y given $\underline{x}$ is

$$h\{[f^{(\lambda)}(\underline{x};\underline{\beta}) + \sigma F^{-1}(p)],\lambda\} \tag{4}$$

and the conditional mean of y is

$$E(y|\underline{x}) = \int_{-\infty}^{a} h\{[f^{(\lambda)}(\underline{x};\beta) + \sigma\epsilon],\lambda\}dF(\epsilon), \tag{5}$$

where $-a\leq x\leq a$ is the support of F. Ruppert and Carroll (1986) discuss estimation of (4) and (5). $E(y|\underline{x})$ is easily estimated by Duan's (1983) "smearing" estimate, which estimates F by the empirical distribution of the residuals; see section 5.

Many data sets we examined have substantial outliers in the untransformed response y, but not in the residuals $y^{(\lambda)}-f^{(\lambda)}(\underline{x};\underline{\beta})$; the transformation has accommodated, or explained, the outlying y's. There is still the danger, however, that a few outliers in y can greatly affect $\hat{x}$ and $\hat{\underline{\beta}}$. Outliers should not be automatically deleted or downweighted, especially when they appear to be part of the normal variation in the response, but it should be standard practice to detect and scrutinize influential cases and when outliers are present to compare the MLE with a robust estimtor. In this paper we propose a diagnostic and a "bounded-influence" estimator which can be used together for detecting influential cases and for robustly estimating λ and $\underline{\beta}$.

Case deletion diagnostics for linear regression are discussed in Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982), and have been extended to the response transformation model (2) by Cook and Wang (1983) and Atkinson (1986). The last two papers approximate the change in $\hat{\lambda}$ as single cases or subsets of cases are deleted. Subset deletion can be unwieldly because of the large number of possible subsets. If influential subsets are to be detected, one needs some strategy to searching for them. Alternatively, one can examine weights from a robust estimator with good breakdown properties.

Bounded-influence regression estimators, so-called because they place a bound on the influence of each observation, have been proposed by Krasker (1980), Hampel (1978), and Krasker and Welsch (1982), and this last paper provides a good overview. Carroll and Ruppert (1985) proposed a bounded influence transformation (BIT) estimator extending the Krasker-Welsch estimator to the response transformation model (2).

In this paper we adapt Atkinson's (1986) diagnostics and the BIT estimator to the "transform both sides" model. The basic technique is to linearize the model (1) by a Taylor approximation at the MLE, and then to apply ordinary regression diagnostics and bounded-influence estimates.

Our methods are designed to be easily implemented on standard software. All our computations were performed on the SAS package using PROC NLIN and rather simple data manipulations in PROC MATRIX and DATA steps. The computations would also be straightforward on other software packages. Our computational techniques can be applied to a bounded-influence estimate for the response-transformation model (2), thus eliminating the need for a lengthy FORTRAN program used in Carroll and Ruppert (1985).

MAXIMUM LIKELIHOOD ESTIMATION

Throughout this paper $y^{(\lambda)}$ is the modified power transformation (3). Under model (1) the log-likelinood is

$$L(\underline{\beta},\lambda,\sigma) = \sum_{i=1}^{N} \log f(\underline{\beta},\lambda,\sigma)$$

where

$$\log f(\underline{\beta},\lambda,\sigma) = -\tfrac{1}{2}\log(2\pi\sigma^2) + (\lambda-1)\log(y_i)$$

$$-1/(2\sigma^2)[y_i{}^{(\lambda)}-f^{(\lambda)}(\underline{x}_i;\underline{\beta})]^2. \tag{6}$$

For fixed $\underline{\beta}$ and λ

$$\hat{\sigma}^2(\underline{\beta},\lambda) = N^{-1}\sum_{i=1}^{N} [y_i^{(\lambda)}-f^{(\lambda)}(\underline{x}_i;\underline{\beta})]^2$$

maximizes $L(\underline{\beta},\lambda,\sigma)$ in σ. Thus, the MLE of $\underline{\beta}$ and λ maximizes

$$L_{max}(\underline{\beta},\lambda) = L(\underline{\beta},\lambda,\hat{\sigma}^2(\underline{\beta},\lambda))$$

$$= -(N/2)\log\{N^{-1}\sum_{i=1}^{N} [(y_i^{(\lambda)}-f^{(\lambda)}(\underline{x}_i;\underline{\beta}))/\dot{y}^{\lambda-1}]^2\} + \text{constant} \tag{7}$$

where $\dot{y}$ is the geometric mean of $y_1,\ldots,y_N$. Therefore, $\hat{\underline{\beta}}$ and $\hat{\lambda}$ minimize

$$\sum_{i=1}^{N} \{y_i^{(\lambda)}-f^{(\lambda)}(\underline{x}_i;\underline{\beta}))/\dot{y}^{\lambda-1}\}^2 \tag{8}$$

Following Box and Cox (1964), $\hat{\underline{\beta}}$ and $\hat{\lambda}$ can be computed as follows. For fixed λ, minimize (8) in $\underline{\beta}$ by ordinary (typically nonlinear) least-squares and call the minimizer $\hat{\underline{\beta}}(\lambda)$. Plot $L_{\max}(\hat{\underline{\beta}}(\lambda),\lambda)$ on a grid and maximize graphically or numerically. This technique is particularly attractive when f is not transformed and $f(\underline{x};\underline{\beta}) = \underline{x}^T\underline{\beta}$ for then (8) can be minimized in $\underline{\beta}$ by linear least-squares. When transforming both sides, the technique is less attractive computationally but it does give the confidence interval

$$\{\lambda:\ L_{\max}(\hat{\underline{\beta}}(\lambda),\lambda) \geq L_{\max}(\hat{\underline{\beta}}(\hat{\lambda}),\hat{\lambda})-\tfrac{1}{2}\chi_1^2(1-\alpha)\}, \tag{9}$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$ quantile of the chi-square distribution with one degree of freedom. Minimizing (8) simultaneously in λ and $\underline{\beta}$ is straightforward with standard nonlinear regression software. One simply fits the dummy variable $D_i \equiv 0$ to the pseudo-model

$$D_i = [y_i^{(\lambda)}-f^{(\lambda)}(\underline{x}_i;\underline{\beta})]/\dot{y}^{\lambda-1} \tag{10}$$

with regression parameter $(\underline{\beta},\lambda)$. Not only is the least-squares estimate of $(\underline{\beta},\lambda)$ the MLE, but for small values of σ and large N estimating the covariance matrix of $\hat{\underline{\beta}}$ using the pseudo regression model (10) is essentially equivalent to inverting the Fisher information for (λ,β,σ), see the appendix.

3. DIAGNOSTICS

Let $(\hat{\lambda},\hat{\underline{\beta}})$ and $(\hat{\lambda}_{(i)},\hat{\underline{\beta}}_{(i)})$ be the MLE's with and without case i, respectively. The changes $\Delta\hat{\lambda}_i = (\hat{\lambda}-\hat{\lambda}_{(i)})$ and $\Delta\hat{\underline{\beta}}_i = (\hat{\underline{\beta}}-\hat{\underline{\beta}}_i)$ are easily interpreted measures of influence, called the sample influence curve (Cook and Weisberg (1982)). Unlike in linear regression, $\Delta\lambda_i$ and $\Delta\hat{\underline{\beta}}_i$ cannot be computed exactly without actually recomputing the MLE with case i deleted. However, $\Delta\lambda_i$ and $\Delta\hat{\underline{\beta}}_i$ can be approximated by applying Atkinson's (1986) "quick estimate" to a linearization of model (1).

To approximate $\Delta\hat{\lambda}_{(i)}$ and $\Delta\hat{\underline{\beta}}_{(i)}$ we linearize the model

$$y^{(\lambda)}/\dot{y}^{\lambda-1} = f^{(\lambda)}(\underline{x};\beta)/\dot{y}^{\lambda-1} + \text{error} \tag{11}$$

about $\hat{\lambda}$, $\hat{\underline{\beta}}$. Let

$$z(\lambda,\underline{\beta}) = [y^{(\lambda)}-f^{(\lambda)}(\underline{x};\underline{\beta})]/\dot{y}^{\lambda-1},$$

$$w(\lambda,\underline{\beta}) = (\partial/\partial\lambda)z(\lambda,\beta),$$

$$u_i(\lambda,\underline{\beta}) = (\partial/\partial\underline{\beta}_i)z(\lambda,\underline{\beta}) = -(\partial/\partial\underline{\beta}_i)f^{(\lambda)}(x;\underline{\beta})/\dot{y}^{\lambda-1},$$

and

$$\underline{u}(\lambda,\underline{\beta}) = (u_i(\lambda,\underline{\beta}),\ldots,u_p(\lambda,\underline{\beta}))^T.$$

Sometimes we will write $z(y,\underline{x};\underline{\beta},\lambda)$ instead of $z(\underline{\beta},\lambda)$ to emphasize the dependence on y and $\underline{x}$. The same holds for $w(\lambda,\underline{\beta})$ and $\underline{u}(\lambda,\underline{\beta})$. Also let $z = z(\hat{\lambda},\hat{\underline{\beta}})$, $w = w(\hat{\lambda},\hat{\underline{\beta}})$, and $\underline{u} = \underline{u}(\hat{\lambda},\hat{\underline{\beta}})$. Then (11) is approximated by

$$z = -(\lambda-\hat{\lambda})w - (\underline{\beta}-\hat{\underline{\beta}})^T\underline{u} + \text{error}. \tag{12}$$

If we fit equation (12) to the full data, then of course $\lambda = \hat{\lambda}$ and $\underline{\beta} = \hat{\underline{\beta}}$. If instead we fit (12) with the ith case deleted, then we obtain Atkinson's (1986) "quick estimate" approximation, which we call $\hat{\lambda}^Q_{(i)}$ and

$\hat{\underline{\beta}}^Q_{(i)}$. Let $\triangle\lambda_i^Q = (\hat{\lambda}-\hat{\lambda}^Q_{(i)})$ and $\triangle\underline{\beta}_i^Q = (\hat{\underline{\beta}}-\hat{\underline{\beta}}^Q_{(i)})$ be the resulting approximations to $\triangle\lambda_i$ and $\triangle\hat{\underline{\beta}}_i$. Because (12) is linear, refitting without case i is easy using standard matrix identities, which have been programmed in many statistical packages.

We used the following computational scheme on SAS. First (8) was minimized by PROC NLIN to obtain $(\hat{\lambda},\hat{\underline{\beta}})$, then z, w, and $\underline{u}$ were generated in a DATA step, and finally the linear model (12) was fit on PROC REG. PROC REG calculates the regression diagnostic DFBETAS_i (Belsley, Kuh, and Welsch 1980), which is a scaled version of $(\triangle\hat{\lambda}_i^Q,\triangle\hat{\underline{\beta}}_i^Q)$. The unscaled $(\triangle\lambda_i^Q,\triangle\hat{\beta}_i^Q)$ is DFBETA_i in the Belsley, Kuh, Welsch nomenclature and is not part of standard SAS output. If we are interested in cases with **relatively** large values of $\triangle\hat{\lambda}_i$ then the scaling is immaterial.

Atkinson's (1986) equation (19) gives a simple formula for calculating $\triangle\hat{\lambda}_i^Q$ alone. We feel, however, that influence for both $\hat{\lambda}$ and $\hat{\underline{\beta}}$ should probably be assessed together and DFBETAS_i is ideal for this. When only the response is transformed, $\hat{\underline{\beta}}$ depends heavily on $\hat{\lambda}$ and it is sensible to estimate $\hat{\lambda}$ first and then to estimate $\hat{\underline{\beta}}$. When transforming both sides, $\hat{\underline{\beta}}$ is usually very stable as $\hat{\lambda}$ is perturbed and $\hat{\lambda}$ and $\hat{\underline{\beta}}$ can be treated simultaneously.

In the example in section 5 and in other examples that we will not report, $\triangle\hat{\lambda}_i$ and $\triangle\hat{\lambda}_i^Q$ were often considerably different, which is surprising since the quick estimate is reasonably accurate when y alone is transformed (Atkinson 1986). The difference is that here $\underline{u}$ depends on $\hat{\lambda}$ and $\hat{\underline{\beta}}$. The approximation $\triangle\hat{\lambda}_i^Q$ does indicate cases with relatively large values of $\triangle\hat{\lambda}_i$, and $\triangle\hat{\lambda}_i^Q$ seems adequate for diagnostic purposes.

To obtain a single measure of joint influence for $(\lambda,\underline{\beta})$ one can compute Cook's D or DFFITS (Belsley, Kuh, and Welsch (1980)) for the psuedo model (12).

4. ROBUST ESTIMATION

A general approach to robust estimation is to minimize asymptotic variance subject to a bound on the gross-error sensitivity. This approach was begun by Hampel (1968, 1974), applied to regression by Hampel (1978), Krasker (1980) and Krasker and Welsch (1982), and used in the response transformation problem by Carroll and Ruppert (1985).

Here we will find an estimator bounding the influence for the parameters λ and $\underline{\beta}$. We will ignore σ, which can be estimated separately with a robust scale functional, e.g. the MAD, applied to the residuals. Let $\ell(\underline{x},y;\lambda,\beta)$ be the score function

$$\ell(\underline{x},y;\lambda,\underline{\beta}) = \tfrac{1}{2}\begin{pmatrix}\partial/\partial\lambda \\ \partial/\partial\underline{\beta}\end{pmatrix} z^2(\lambda,\underline{\beta}) \tag{13}$$

$$= z(\lambda,\underline{\beta})\begin{pmatrix}w(\lambda,\underline{\beta}) \\ \underline{u}(\lambda,\underline{\beta})\end{pmatrix}.$$

Since the MLE minimizes (8) it solves

$$\sum_{i=1}^{N} \ell(\underline{x}_i,y_i;\lambda,\underline{\beta}) = 0,$$

at least when λ and $\underline{\beta}$ are unconstrained and $f(\underline{x};\underline{\beta})$ is a smooth function of $\underline{\beta}$. The MLE is highly sensitive to cases with large values of $z(\hat{\lambda},\hat{\underline{\beta}})$, $w(\hat{\lambda},\hat{\underline{\beta}})$, or $\underline{u}(\hat{\lambda},\hat{\underline{\beta}})$ corresponding to response outliers, high leverage points for λ, and high leverage points for $\underline{\beta}$, respectively.

A robust bounded-influence estimator $(\tilde{\lambda},\tilde{\underline{\beta}})$ is found by solving

$$\sum_{i=1}^{N} W(y_i,\underline{x}_i;\tilde{\lambda},\tilde{\underline{\beta}})\ell(y_i,\underline{x}_i;\tilde{\lambda},\tilde{\underline{\beta}}) = 0$$

where W is a scalar weight function such that $W\ell$ is bounded. The optimal choice of W was first studied by Hampel (1968, 1974) for general univariate parametric families. When choosing W, asymptotic efficiency measured by the covariance of $(\tilde{\lambda},\underline{\beta})$ must be balanced against robustness measured by the norm of $W\ell$ For a multivariate parameter such as $(\lambda,\underline{\beta})$, this balancing raises philosophical questions since there are many ways of comparing covariance matrices or of norming vector functions. The approach we take generalizes the Krasker-Welsch (1982) bounded-influence regression estimates. Whether the Krasker-Welsch estimator is optimal in any meaningful sense is an open question (Ruppert 1985), but it seems quite satisfactory in practice.

Let ℓ be the gradient of log $f(\underline{\beta},\lambda,\sigma)$ with respect to $(\underline{\beta},\lambda)$. For any weighting function W, the influence function evaluated at $(y_i,\underline{x}_i)$ will be defined as

$$IF(y,\underline{x};\lambda,\underline{\beta}) = B^{-1}W(y,\underline{x};\lambda,\underline{\beta})\ell(y,\underline{x};\lambda,\underline{\beta})$$

where

$$B = N^{-1}\sum_{i=1}^{N} E_y\{W(y,x_i,\lambda,\underline{\beta})\ell(y,\underline{x}_i;\lambda,\underline{\beta})\ell^T(y,\underline{x}_i;\lambda,\underline{\beta})\}.$$

This definition of IF coincides wiith the usual definition when the x's are i.i.d. for some H, and the averaging over $\{x_1,\ldots,x_N\}$ in the definition of B is replaced by expectation with respect to H. Our definition is appropriate for fixed or random x's. In the definition of B on page 5 of Carroll and Ruppert (1985), W is incorrectly squared. In that article, but not here, $\ell = 1$. The asymptotic covariance matrix of

$(\hat{\lambda},\hat{\underline{\beta}})$ is $V = B^{-1}A(B^{-1})^T$ where

$$A = N^{-1}\sum_{i=1}^{N} E_y\{W^2(y,\underline{x}_i,\lambda,\underline{\beta})\psi(y,\underline{x}_i;\lambda,\underline{\beta})\psi^T(y,\underline{x}_i;\lambda,\underline{\beta})\}.$$

An intuitively reasonable way to norm $W\psi$ is to use the asymptotic covariance; see Krasker and Welsch (1982) for further motivation and discussion. The resultant measure of influence, the so-called self-standardized gross-error sensitivity is

$$\gamma_2 = \max_i\ [IF(y_i,\underline{x}_i;\lambda,\hat{\underline{\beta}})^T V^{-1} IF((y_i,\underline{x}_i,\lambda,\underline{\beta})]$$

$$= \max_i\ \{[\psi(y_i,\underline{x}_i,\lambda,\underline{\beta})^T A^{-1}\psi(y_i,\underline{x}_i,\lambda,\underline{\beta})]^{\frac{1}{2}} W^2(y_i,\underline{x}_i;\lambda,\underline{\beta})\}.$$

Note that W^2 has been incorrectly omitted from the last term in equation (15) of Carroll and Ruppert (1985). γ_2 must be at least $(p+1)^{\frac{1}{2}}$. From experience with other problems we suggest bounding γ_2 by $a(p+1)^{\frac{1}{2}}$, where "a" is between 1.1 and 1.5, and a = 1.2 or 1.3 has generally been satisfactory. To bound γ_2 by $a(p+1)^{\frac{1}{2}}$, we use the weighting function

$$W(y,x;\lambda,\underline{\beta}) = \min\{1,a(p+1)^{\frac{1}{2}}[\psi(y,\underline{x};\lambda,\underline{\beta})^T A^{-1}\psi(y,x;x,\underline{\beta})]^{-\frac{1}{2}}\}.$$

Here A is defined implicitly since it depends upon W and vice versa. In practice, λ, $\underline{\beta}$, and A are estimated iteratively.

We used a simple iterative scheme:

(1) Fix a>1. Let C be the total number of cycles. Set c=1. Let $\hat{\lambda}_p$ and $\hat{\underline{\beta}}_p$ be preliminary estimates, probably the MLEs. Set $W_i \equiv 1$.

(2) Define

$$\hat{A} = N^{-1}\sum_{i=1}^{N} W_i^2 \psi(y_i,\underline{x}_i,\hat{\lambda}_p,\hat{\underline{\beta}}_p)\psi^T(y_i,\underline{x}_i,\hat{\lambda}_p,\hat{\underline{\beta}}_p).$$

In ψ use the weighted geometric mean $\dot{y} = \exp(\sum w_i \log y_i/\sum w_i)$.

(3) Update the weights:

$$W_i = \min\{1,\ a(p+1)^{\frac{1}{2}}[\psi(y_i,\underline{x}_i,\hat{\lambda}_p,\hat{\underline{\beta}}_p)^T\hat{A}^{-1}\psi(y_i,\underline{x}_i;\hat{\lambda}_p,\hat{\underline{\beta}}_p)]^{-\frac{1}{2}}\}.$$

(4) Solve

$$\sum_{i=1}^{N} W_i\psi(y_i,\underline{x}_i;\hat{\lambda},\hat{\underline{\beta}}) = 0$$

(5) If $c<C$, set $\hat{\lambda}_p = \hat{\lambda}$ and $\hat{\underline{\beta}}_p = \hat{\underline{\beta}}$, $c=c+1$ and return to (2). If $c=C$ then stop.

In the examples discussed in section 5 and other examples that we will not report, $\hat{\lambda}$ and $\hat{\underline{\beta}}$ stabilized at C=2. Therefore, we recommend C=2, or perhaps C=3 for small N or data sets with extremely influential points. In fact C=1 seems adequate, at least for diagnostic purposes. We calculated step (3) with a short program in PROC MATRIX and step (4) was performed in PROC NLIN. PROC MATRIX is needed only to invert A, and the program should be easily modified when PROC MATRIX is replaced by an interactive matrix language. Undoubtedly, the computations would also be easy on other packages. We will call the final estimate the BITBS.

This iterative method can also be used for the response transformation model. Instead of using (3) one sets

$$\psi(\underline{x},y;\lambda,\underline{\beta}) = \begin{pmatrix}\partial/\partial\lambda\\ \partial/\partial\underline{\beta}\end{pmatrix}\{[y^{(\lambda)}-f(\underline{x};\underline{\beta})]/\dot{y}^{\lambda-1}\}^2.$$

It should be noted that this approach differs from the BIT estimate of

Carroll and Ruppert (1985), since the BIT estimates σ simultaneously. Because the likelihood scores for $\underline{\beta}$ and σ are, respectively, linear and quadratic in the residual, the joint bounded-influence estimator of λ,$\underline{\beta}$, and σ behaves as a rescending "psi-function", e.g. a Hampel M-estimator (Hampel 1978). Therefore, the influence on λ and $\underline{\beta}$ of extreme response outliers approaches zero when the influence for σ is simultaneously bounded.

5. AN EXAMPLE

When managing a fish stock, one must model the relationship between the annual spawning stock size and the eventual production of new catchable-sized fish (returns or recruits) from the spawning. Ricker and Smith (1975) give numbers of spawners (S) and returns (R) from 1940 until 1967 for the Skeena River sockeye salmon stock. Using some simple assumptions about factors influencing the survival of juvenile fish, Ricker (1954) derived the theoretical model

$$R = \beta_1 S \exp(\beta_2 S) = f(S;\underline{\beta}) \tag{14}$$

relating R and S. Other models have been proposed, e.g. by Beverton and Holt (1957). However, the Ricker model appears adequate and, in particular, gives almost the same fit for this stock as the Beverton-Holt model.

From figure 1, a plot of R against S, it is clear that R is highly variable and heteroscedastic, with the variance of R increasing with its mean. Several cases appear somewhat outlying, in particular #5, #19, and #25. The model (14) was linearized about the MLE to form the pseudo model (12) and the square root of Cook's D was plotted against case number; see figure 2. Case #5 and especially case #12 stand out. In figure 1 case #12 is somewhat masked by the heteroscedasticity since the residual on the original scale $[R_{12}-f(S_{12};\hat{\underline{\beta}})]$ is relatively small, but after transformation by the MLE $\hat{\lambda} = .314$ the residual $[R_{12}^{(\hat{\lambda})}-f^{(\hat{\lambda})}(S;\hat{\underline{\beta}})]$ is substantially larger though still not excessive. Case #12 is an extremely high leverage point, and its Hat matrix diagonal according to model (12) is $h_{12} = 0.685$. An h value exceeding $2p/N = 6/28 = .214$ is considered high by Hoaglin and Welsch (1978). Since $h_5 = .23$, case #5 is also a leverage point by this criterion. In figure 3, the residuals $[R_i^{(\hat{\lambda})}-f^{(\hat{\lambda})}(S_i,\hat{\underline{\beta}})] = \hat{e}_i$ are plotted against S and though #12 stands out, it does not seem extremely outlying until one accounts for leverage. To compensate for leverage, Belsley, Kuh, and Welsch (1980) suggest standardizing $\hat{e}_i$ by its estimated standard error to produce

$$\mathrm{RSTUDENT}_i = \hat{e}_i/S(i)(1-h_i)^{\frac{1}{2}},$$

where S(i) is the root mean square error without case i. For this data set $\mathrm{RSTUDENT}_{12} = -4.40$!

In table 1 we present influence diagnostics applied to model (12), the exact change $\Delta\hat{\lambda}_i$, the quick estimate $\Delta\hat{\lambda}_i^Q$, and another approximation $\Delta\hat{\lambda}_i^N$. It is evident that $\Delta\hat{\lambda}_i^Q$ is not always close to $\Delta\hat{\lambda}_i$, and there are at least two possible causes of inaccuracy: (i) $\Delta\hat{\lambda}_i^Q$ uses a linearization of the parameters and (ii) in model (12) $\dot{y}$ is held fixed rather than readjusted as cases are deleted. To isolate the effects of cause (ii), we experimented with a different approximation. We computed the nonlinear least squares estimate $\hat{\lambda}_{(i)}^N$ minimizing (8) when the $\dot{y}$ is calculated from the full data but the sum of squares in (8) is over $j \neq i$. Then we set $\Delta\hat{\lambda}_i^N = (\hat{\lambda}-\hat{\lambda}_{(i)}^N)$. Of course, $\Delta\hat{\lambda}_i^N$ is as difficult to compute as $\Delta\hat{\lambda}_i$ itself and is not of interest as a practical approximation; we have calculated $\Delta\hat{\lambda}_i^N$ just to learn why $\Delta\hat{\lambda}_i^Q$ is inaccurate. In table 1, $|\Delta\hat{\lambda}_i|$ is large for i=5 and 12. In both cases, $\Delta\hat{\lambda}_i^N$ approximates $\Delta\hat{\lambda}_i$ much better than $\Delta\hat{\lambda}_i^Q$ which suggests that cause (i) is the primary problem. It is clear, though, that small changes in $\dot{y}$ from deleting an outlying y_i can have a notable effect on $\hat{\lambda}$. We have compared $\Delta\lambda_i$ and $\Delta\lambda_i^Q$ for several other data sets, the kinetics data of Carr (1960) analyzed in Box and Hill (1974) and Carroll and Ruppert (1984), the Atlantic menhaden spawner-recruit data in Carroll and Ruppert (1985), and the "population A" spawner-recruit data in Ruppert and Carroll (1986). In all cases $\Delta\lambda_i^Q$, though not an accurate approximation, did indicate large values of $|\Delta\hat{\lambda}_i|$, and $\Delta\lambda_i^Q$ appears to be an effective diagnostic of high influence.

We computed the BITBS estimate with bound a=1.2 and C=3 cycles. In table 2, $\hat{\lambda}$, $\hat{\beta}_1$,$\hat{\beta}_2$ and non-unit w_i are given for the MLE and each iteration of BITBS. The changes in the estimates and weights from the first to the second iteration are small, and one iteration seems adequate for most practical purposes; certainly two iterations suffice. Although the BITBS estimate severely downweights case #12, the estimate of λ only changes from .31 to .13 in two iterations of BITBS, while the MLE becomes $\hat{\lambda}$ = -.2 if #12 is deleted. For these data, the BITBS detects the influential points and reduces, but does not eliminate, their influence. Case #12 was the year 1951 when a rock slide severely reduced recruitment (Ricker and Smith 1975). To model recruitment under normal conditions one would delete case #12 and refit. Without #12, the MLE and the BITBS estimate are similar (table 3), and one would probably use the MLE. The bulk of the data indicate that recruitment is highly heteroscedastic and a severe transformation ($\hat{\lambda}$ = -.2) is needed to induce homoscedastic errors. Because the anomalous #12 occurs where S is small, it indicates less heteroscedasticity and a more moderate transformation ($\hat{\lambda}$ = .3) is used if #12 is not deleted. Since case #12 is not from the target population of normal spawning years, it seems safe to delete it.This data set is an example where one might consider a robust estimator that gives essentially zero influence to extreme outliers, e.g. a generalization to transformation models of a redescending-psi M-estimator. We normally prefer using a redescending M-estimator to rejecting outliers, since an M-estimator has a known large sample distribution. The effects of outlier rejection methods upon the MLE are not well understood, even asymptotically. Here we distinguish between rejecting outliers based on some statistical criterion, say a hypothesis test, and deleting observations known to come from a non-target population.

Although the MLE and the BITBS are similar, when #12 is deleted, two cases, #4 and #5, are substantially downweighted by the BITBS. Case #4 had been masked when #12 was included, but #5 was already influential in the presence of #12. One might explore further deletions though without further information we feel that #4 and #5 should be retained in the final analysis. If in addition to #12, case #4, case #5, or both are deleted, then the MLE changes noticeably; see Table 4. The deletion of #5 and the deletion of #4 affect the MLE in somewhat opposite directions, though deleting #4 affects the MLE more severely. The BITBS downweights #4 somewhat more than #5, so effects of downweighting #4 and #5 tend to cancel.

We do not view the "transform both sides" method chiefly as choosing a new scale for analyzing the response, but rather as modeling the conditional distribution of y on the original scale. By "original" scale, we mean the scale of primary scientific interest, usually also the scale on which direct measurements have been made. The model

$$y^{(\lambda)} = f^{(\lambda)}(\underline{x};\underline{\beta})+\epsilon$$

leads to

$$y = y(\epsilon) = (f^{\lambda}(\underline{x};\underline{\beta})+\lambda\epsilon)^{1/\lambda} \qquad \lambda \neq 0,$$
$$= \exp(\log f(\underline{x};\underline{\beta}) + \lambda\epsilon) \qquad \lambda = 0.$$

We are assuming that ϵ is approximately normal and in particular approximately symmetric, and this last point suggests that the conditional median of y given $\underline{x}$ can be estimated by

$$\hat{m}(y|\underline{x}) = f(\underline{x};\hat{\underline{\beta}}).$$

The conditional mean is easily estimated by the "smearing" estimate of Duan (1983). Let r_i by the i-th residual

$$r_i = y_i^{\lambda} - f^{\lambda}(\underline{x}_i, \hat{\underline{\beta}}) = \lambda(y_i^{(\lambda)} - f^{(\lambda)}(\underline{x}_i, \hat{\underline{\beta}})).$$

Then the smearing estimate is

$$\hat{E}(y|\underline{x}) = N^{-1} \sum_{i=1}^{N} \{f^{\hat{\lambda}}(\underline{x};\hat{\underline{\beta}}) + r_i\}^{1/\hat{\lambda}}, \tag{15}$$

with obvious changes if $\lambda = 0$.

If λ is 0 or if λ^{-1} is an integer, then $E(y|\underline{x})$ can be estimated by methods of Miller (1984). Miller's estimators are particularly simple when λ is in the set $\{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$. It is a common practice to round $\hat{\lambda}$ to a value in this set, in which case Miller's (1984) estimate is applicable. We do not necessarily advocate rounding $\hat{\lambda}$ especially since there is some theoretical evidence against this practice (Carroll 1982), but when the rounded value is very plausible according to a hypothesis test, then the rounding should have little effect on subsequent inference. The common rationale for rounding $\hat{\lambda}$, to make the transformation easily interpretable, is less compelling when one views λ as part of a model for y on the original scale. In the present example, $R^{.3}$ or $R^{-.2}$ is admittedly of little direct biological and economic interest, but the same is true of, say, log(R).

In figure 1, $\hat{m}(R|S)$ and $\hat{E}(R|S)$, both calculated without #12, are plotted. $E(R|S)$ was also estimated by fixing $\lambda = 0$, re-estimating $\hat{\underline{\beta}}$ and $\hat{\sigma}$, and using Miller's (1984) estimate:

$$f(\underline{x};\hat{\underline{\beta}})e^{\hat{\sigma}^2/2}.$$

The smearing estimate and Miller's estimate are so close that they would be barely distinguishable had Miller's estimate be included in figure 1.

To see the influence of case #12 on $\hat{m}(R|S)$ and $\hat{E}(R|S)$, these estimates were calculated both with and without case #12. When #12 is deleted then not only are $\hat{\lambda}$ and $\hat{\underline{\beta}}$ set equal to the MLE without #12, but also the averaging in (15) is over $i \neq 12$.

The changes in the estimated median and mean caused by deletion of case #12 are graphed in figure 4. As might be expected, deleting #12 caused the estimated median and mean recruitment to increase for small S, especially for S near 300-400. The most dramatic change when deleting #12 is a decrease in estimated median and mean recruitment for large S. This decrease is largely brought about by the decrease in $\hat{\beta}_2$, since β_2 controls the shape of the Ricker curve.

The "transform both sides" model is certainly not the only model that would be appropriate for this example. Since R is heteroscedastic but not greatly skewed one should consider heteroscedastic models such as

$$R = f(S;\underline{\beta}) + \sigma S^{\alpha}\epsilon \tag{16}$$

or

$$R = f(S;\underline{\beta}) + \sigma f^{\alpha}(S;\underline{\beta})\epsilon,$$

where the variance of R is proportional to a power of S or of f. In Ruppert and Carroll (1986), the model

$$R^{(\lambda)} = f^{(\lambda)}(S;\underline{\beta}) + \sigma S^{\alpha}\epsilon \tag{17}$$

was fit to the Skeena River data, with case #12 included. The MLE was $\hat{\lambda} = .75$ and $\hat{\alpha} = .5$. However, both H_o: $\alpha = 0$ and H_o: $\lambda = 1$ are accepted by likelihood ratio tests at level .10, so models (14) and (16) both appear reasonable for these data, though a re-analysis without case #12 would be of interest. We plan to study diagnostics and robust estimation for model (17) in the future.

6. SUMMARY

When a response y is thought to fit a model $f(\underline{x};\underline{\beta})$, but y is heteroscedastic and/or nonnormally distributed, then y and $f(\underline{x};\underline{\beta})$ can be transformed in the same manner to induce approximately homoscedastic, normal errors while retaining the model $f(\underline{x};\underline{\beta})$ for the conditional median of y. Often outliers in the original response are accommodated by transformation; that is, the outliers are seen to be the result of the skewness or heteroscedasticity in the untransformed data.

In some situations, an outlier will indicate a substantially different transformation than that fitting the bulk of the data. In our example with 28 observations, case #12 is a response outlier associated with a small value of the conditional median (and mean) response. Therefore, case #12 counter-indicates the severe heteroscedasticity in the rest of the data, and deleting #12 changes the estimated power transformation from $\hat{\lambda} = .3$ to $\hat{\lambda} = -.2$.

Influential cases should be detected and scrutinized as a matter of standard good statistical practice. In some situations, such as with case #12 in our example, there are good reasons for removing an influential case. In other cases, the appropriate treatment of the outliers will be less clear-cut.

In this paper, we propose an approximation to the sample influence curve. Although the approximation is not highly accurate it is an effective diagnostic for influence cases. We also propose a bounded influence estimator, which can be used to pinpoint influencial cases, or to accommodate them, or both. The diagnostic and the robust estimator can both be computed with standard software.

APPENDIX

There are at least three methods of estimating the covariance matrix of $(\hat{\underline{\beta}},\hat{\lambda})$: (i) Evaluate the Hessian of $-L(\underline{\beta},\lambda,\sigma)$ at $(\hat{\underline{\beta}},\hat{\lambda},\hat{\sigma})$. This is the observed Fisher information matrix, I. Use the $(p+1)\times(p+1)$ upper-left submatrix of I^{-1}. (ii) Let I^{*} be the Hessian of $-L_{max}(\underline{\beta},\lambda)$ evaluated at $(\hat{\underline{\beta}},\hat{\lambda})$. Use $(I^{*})^{-1}$. (iii) Use the estimate from model (10) treated as a nonlinear reoression problem. Also there is a fourth method which only estimates the covariance matrix of $\hat{\underline{\beta}}$. Suppose we followed the Box-Cox method of estimation described in section 2. Then we have obtained $\hat{\lambda}$ and we have estimated $\underline{\beta}$ by $\hat{\underline{\beta}}(\hat{\lambda})$, which is the least squares estimate when $y^{(\hat{\lambda})}$ is fit to $f^{(\hat{\lambda})}(\underline{x};\underline{\beta})$ with λ fixed at $\hat{\lambda}$. This fit also gives an estimated covariance matrix for $\hat{\underline{\beta}}$, which we will call the method (iv) estimate.

Methods (iii) and (iv) are the easiest to use since they can be implemented on standard nonlinear regression software. Method (i) is justified by the well-known large sample theory of maximum likelihood estimation. Method (ii) ignores σ and treats $L_{max}(\underline{\beta},\lambda)$ as a likelihood for $(\underline{\beta},\lambda)$.

In theorem A.2 below, we show that methods (i) and (ii) are identical, not just for the transformation problem under study but in general for parametric estimation where a parameter is eliminated by maximizing the likelihood over that parameter.

In general, methods (i) and (ii) are not equivalent to (iii) even as $N \to \infty$, but by theorem A.1 below all three estimates of $\text{var}(\hat{\underline{\beta}})$ are the same in the limit as $N \to \infty$ and $\sigma \to 0$. Bickel and Doksum (1981) and Carroll and Ruppert (1984) have let $N \to \infty$ and $\sigma \to 0$ simultaneously to provide a simple asymptotic theory for transformations, since the usual $N \to \infty$ and σ fixed theory is complicated. "Small σ" asymptotics have often proved to be good approximations to finite sample results when checked against Monte Carlo. Moreover, in many data sets, especially from engineering and the physical sciences, σ does seem small in the sense that the model fits the data very well.

In Carroll and Ruppert (1984) we show that methods (i) and (iv) are equivalent estimates of var($\hat{\underline{\beta}}$) as $N \to \infty$ and $\sigma \to 0$.

In summary, methods (i) and (ii) are identical, except of course that method (ii) does not estimate var($\hat{\sigma}$) or the covariance of $\hat{\sigma}$ with $\hat{\underline{\beta}}$ and $\hat{\lambda}$. All four methods give asymptotically equivalent estimates of var($\hat{\underline{\beta}}$) as $N \to \infty$ and $\sigma \to 0$. It does not appear that method (iii) correctly estimates var($\hat{\lambda}$) or cov($\hat{\lambda},\hat{\underline{\beta}}$). The confidence interval (9) for $\hat{\lambda}$ can be used in place of a standard error for $\hat{\lambda}$, but a δ-method standard error of $\hat{E}(y|\underline{x})$ or $\hat{m}(y|\underline{x})$ will require an estimate of cov($\hat{\lambda},\hat{\underline{\beta}}$) and var($\hat{\lambda}$).

Programming method (ii) by computing the analytic second derivative matrix of $L_{\max}$ is somewhat a bother, but the gradient of $L_{\max}$ is easily programmed and can be differentiated numerically. Since

$$L_{\max}(\underline{\beta},\lambda) = -(N/2)\log\hat{\sigma}^2(\underline{\beta},\lambda)$$

where

$$\hat{\sigma}^2(\underline{\beta},\lambda) = N^{-1}\sum_{i=1}^{N} z^2(y_i,\underline{x}_i;\underline{\beta},\lambda)$$

and since the gradient of $\hat{\sigma}^2$ at $(\hat{\underline{\beta}},\hat{\lambda})$ is zero, the Hessian of $L_{\max}$ at $(\hat{\underline{\beta}},\hat{\lambda})$ is

$$L_{\max}(\hat{\underline{\beta}},\hat{\lambda}) = -\hat{\sigma}^{-2}\sum_{i=1}^{N}\begin{bmatrix}(\partial/\partial\underline{\beta})^T(z\underline{u}) & (\partial/\partial\lambda)(z\underline{u})\\ (\partial/\partial\underline{\beta})^T(zw) & (\partial/\partial\lambda)(zw)\end{bmatrix}$$

where all quantities on the right hand side are evaluated at $(\hat{\underline{\beta}},\hat{\lambda})$. It is not difficult to numerically compute the derivatives of $(z\underline{u})$ and (zw) with respect to $\underline{\beta}$ and λ.

In table A.1 we compare the method (ii), (iii) and (iv) standard errors for the Skeena River data with case #12 deleted. The three methods produce similar standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$. The standard error of $\hat{\lambda}$ by method (iii) seems substantially inflated.

<u>Theorem A.1</u> Suppose that $\underline{x}_1, \underline{x}_2, \ldots$ are i.i.d. Then as $N \to \infty$ and $\sigma \to 0$, methods (i) and (iii) of estimating the covariance matrix of $\hat{\underline{\beta}}$ are asymptotically equivalent.

<u>Sketch of proof</u>: Define:

$$\underline{\Sigma}_{11} = \sum_{i=1}^{N} \underline{u}(y_i, \underline{x}_i; \hat{\underline{\beta}}, \hat{\lambda}) \underline{u}^T(y_i, \underline{x}_i; \hat{\underline{\beta}}, \hat{\lambda})$$

$$\underline{\Sigma}_{12} = \sum_{i=1}^{N} \underline{u}(y_i, \underline{x}_i; \hat{\underline{\beta}}, \hat{\lambda}) w(y_i, \underline{x}_i; \hat{\underline{\beta}}, \hat{\lambda})$$

$$\underline{\Sigma}_{22} = \sum_{i=1}^{N} w^2(y_i, \underline{x}_i; \hat{\underline{\beta}}, \hat{\lambda})$$

and

$$\underline{\Sigma} = \begin{bmatrix} \underline{\Sigma}_{11} & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{12}^T & \underline{\Sigma}_{22} \end{bmatrix}.$$

Then the estimated covariance matrix of $(\hat{\underline{\beta}}, \hat{\lambda})$ by method (iii) is $s^2 \underline{\Sigma}^{-1}$ where s^2 is the mean square error. Now as $N \to \infty$

$$N^{-1}\mathcal{I}_{11} \to S := E\underline{u}(y_1,\underline{x}_1;\underline{\beta},\lambda)\underline{u}^T(y_1,\underline{x}_1,\underline{\beta},\lambda),$$

and by Taylor expansions

$$(N\sigma)^{-1}\mathcal{I}_{12} \to 0$$

and

$$(N\sigma^2)^{-1}\mathcal{I}_{22} \to D := E[(\partial/\partial\epsilon_1 w(y_1,\underline{x}_i,\underline{\beta},\lambda)|_{\epsilon_1=0})\epsilon_1]^2$$

as N → ∞ and $\sigma \to 0$. (Note that y_1 is a function of ϵ_1; see equation (1).) Therefore, by method (iii) the estimate of $\mathrm{var}((N^{\frac{1}{2}}/\sigma)\hat{\underline{\beta}}, N^{\frac{1}{2}}\hat{\lambda})$ converges to

$$\begin{bmatrix} S^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix}.$$

By theorem 1 of Carroll and Ruppert (1984) method (i) has the same asymptotic estimate of $\mathrm{var}((N^{\frac{1}{2}}/\sigma)\hat{\underline{\beta}})$ but a different estimate of $\mathrm{var}(N^{\frac{1}{2}}\hat{\lambda})$.

<u>Note</u>: Some type of regularity conditions on $\{\underline{x}_i\}$ are needed for the asymptotics to hold. The assumption that $\{\underline{x}_i\}$ are i.i.d. is convenient but other assumptions could be used instead. For a rigorous result, an appropriate regularity condition on f would also be needed.

Theorem 4.2: Let $L(\underline{\theta},\sigma)$, $\underline{\theta}\in R^k$ and $\sigma\in R^q$, be a real-valued function. Let $L_{\underline{\theta}}$ and L_{σ} be the first partial derivatives and let $L_{\underline{\theta}\underline{\theta}}$, $L_{\underline{\theta}\sigma}$, and

$L_{\sigma\sigma}$ be the second partial derivatives of L. For each $\underline{\theta}$ suppose $\sigma(\underline{\theta})$ satisfies

$$L(\theta,\sigma(\theta)) = \sup_{\sigma} L(\underline{\theta},\sigma). \tag{A.1}$$

Then

$$\{(\partial^{2}/\partial\underline{\theta}\partial\underline{\theta})L(\underline{\theta},\sigma(\underline{\theta}))\}^{-1}$$

is the upper left k×k submatrix of

$$\begin{bmatrix} L_{\underline{\theta\theta}}(\underline{\theta},\sigma(\theta)) & L_{\underline{\theta}\sigma}(\underline{\theta},\sigma(\underline{\theta})) \\ L^{T}_{\underline{\theta}\sigma}(\underline{\theta},\sigma(\underline{\theta})) & L_{\sigma\sigma}(\underline{\theta},\sigma(\underline{\theta})) \end{bmatrix}^{-1} \tag{A.2}$$

<u>Proof</u>: It is enough to prove the theorem when q=1, for then the general case follows by induction. By (A.1)

$$L_{\sigma}(\underline{\theta},\sigma(\underline{\theta})) \equiv 0,$$

so that

$$0 = (\partial/\partial\underline{\theta})L_{\sigma}(\underline{\theta},\sigma(\underline{\theta})) = L_{\underline{\theta}\sigma}(\underline{\theta},\sigma(\underline{\theta})) + L_{\sigma\sigma}(\underline{\theta},\sigma(\underline{\theta}))(\partial\sigma(\underline{\theta})/\partial\underline{\theta}).$$

Therefore

$$\partial\sigma(\underline{\theta})/\partial\underline{\theta} = -L_{\underline{\theta}\sigma}(\underline{\theta},\sigma(\underline{\theta}))/L_{\sigma\sigma}(\underline{\theta},\sigma(\underline{\theta})). \tag{A.3}$$

Next

$$(\partial^{2}/\partial\underline{\theta}\partial\underline{\theta})L(\theta,\sigma(\underline{\theta})) = L_{\underline{\theta\theta}}+ L_{\theta\sigma}(\partial\sigma/\partial\underline{\theta})^{T} + (\partial\sigma/\partial\underline{\theta})L^{T}_{\theta\sigma}$$

$$+ (\partial\sigma/\partial\underline{\theta})(\partial\sigma/\partial\underline{\theta})^{T}L_{\sigma\sigma} \tag{A.4}$$

where all terms on the right-hand side of (A.4) are evaluated at $\underline{\theta},\sigma(\underline{\theta})$. Substituting (A.3) into (A.4) we have

$$(\partial^2/\partial\underline{\theta}\partial\underline{\theta})L(\underline{\theta},\sigma(\underline{\theta})) = L_{\underline{\theta\theta}}-L_{\underline{\theta}\sigma}L^T_{\underline{\theta}\sigma}/L_{\sigma\sigma}.$$

Using the identity, $(A^{-1}+UV^T)^{-1} = A^{-1}-A^{-1}UV^TA^{-1}/(1+V^TA^{-1}U)$ if $A\in R^{k\times k}$ and $U, V \in R^k$ (see problem 2.8, page 33 of Rao 1973) we have

$$\{(\partial^2/\partial\underline{\theta}\partial\underline{\theta})L(\underline{\theta},\sigma(\underline{\theta}))\}^{-1}$$

$$= L^{-1}_{\underline{\theta\theta}} + (L^{-1}_{\underline{\theta\theta}}L_{\underline{\theta}\sigma}L_{\underline{\theta}\sigma}L_{\underline{\theta\theta}}^{-1})/(L_{\sigma\sigma}-L^T_{\underline{\theta}\sigma}L_{\underline{\theta\theta}}^{-1}L_{\underline{\theta}\sigma}). \tag{A.5}$$

By another identity (see problem 2.7, page 33 of Rao 1973), (A.5) is the $k\times k$ upper left submatrix of (A.2).

Table 1

Diagnostics for the Skeena River sockeye salmon data.

Diagnostics	Case Number			
	5	12	19	25
Residual	917	-939	-882	-922
RSTUDENT	2.25	-4.40	-1.93	-2.04
Hat diagonal	.23	.68	.08	.08
Cook's D	.43	8.09	.09	.11
DFFITS	1.23	-6.49	-.55	-.62
DFBETAS-$\hat{\beta}_1$	-.46	-.71	.13	.19
DFBETAS-$\hat{\beta}_2$	.55	1.38	-.33	-.41
DFBETAS-$\hat{\lambda}$	-1.01	6.06	-.11	-.11
$\Delta\hat{\lambda}_i^Q$	-.31	1.56	-.04	-.04
$\Delta\hat{\lambda}_i^N$	-.17	.77	-.01	-.01
$\Delta\hat{\lambda}_i$	-.10	.51	-.03	-.03

Table 2

Maximum likelihood and bounded-influence estimates for the Skeena River data. No cases deleted.

	C=0 (MLE)	C=1	C=2	C=3
$\hat{\beta}_1$	3.295	3.590	3.619	3.622
$\hat{\beta}_2$	$-6.9998x10^{-4}$	$-8.307x10^{-4}$	$-8.49x10^{-4}$	$-8.50x10^{-4}$
$\hat{\lambda}$	.3141	.1921	.1329	.1138
w_5	1.0	.448	.579	.647
w_6	1.0	.931	1.0	1.0
w_{12}	1.0	.253	.188	.172
w_{19}	1.0	.811	.857	.874
w_{25}	1.0	.733	.776	.790

Table 3

Maximum likelihood and bounded-influence estimates for the Skeena River data. Case #12 deleted.

	C=0 (MLE)	C=1	C=2
$\hat{\beta}_1$	3.78	3.98	3.89
$\hat{\beta}_2$	$-9.54\text{x}10^{-4}$	$-10.2\text{x}10^{-4}$	$-9.93\text{x}10^{-4}$
$\hat{\lambda}$	-.199	-.254	-.235
w_4	1.0	.377	.575
w_5	1.0	.448	.753
w_6	1.0	1.0	.946
w_9	1.0	1.0	.954
w_{12}	This case is deleted		
w_{18}	1.0	1.0	.904
w_{19}	1.0	.781	.860
w_{25}	1.0	.703	.846

Table 4

Maximum likelihood estimation for the SkeenaRiver data with selected cases removed

	Cases Removed			
	#12	#4, #12	#5, #12	#4, #5, #12
$\hat{\beta}_1$	3.78	4.20	3.89	4.30
$\hat{\beta}_2$	-9.54×10^{-4}	-11.2×10^{-4}	-10.5×10^{-4}	-12.1×10^{-4}
$\hat{\lambda}$	-.199	-.428	-.126	-.392

Table A.1

Estimated standard errors for the Skeena River sockeye salmon data without case #12

Method	s.e($\hat{\beta}_1$)	s.e.($\hat{\beta}_2$)	s.e.($\hat{\lambda}$)
(ii)	0.698	3.17×10^{-4}	0.369
(iii)	0.711	3.33×10^{-4}	0.624
(iv)	0.694	3.06×10^{-4}	---

LIST OF FIGURES

Figure 1

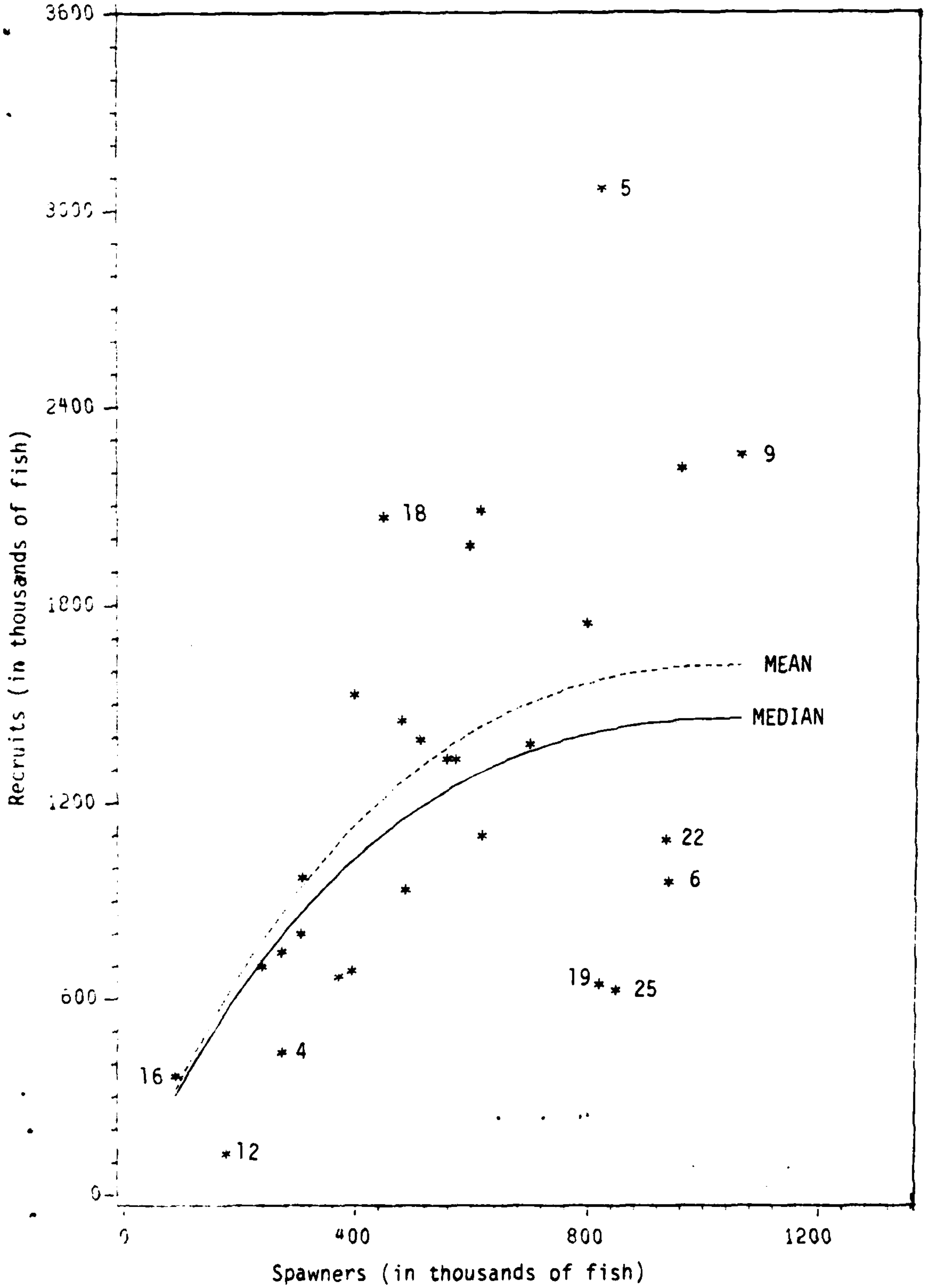
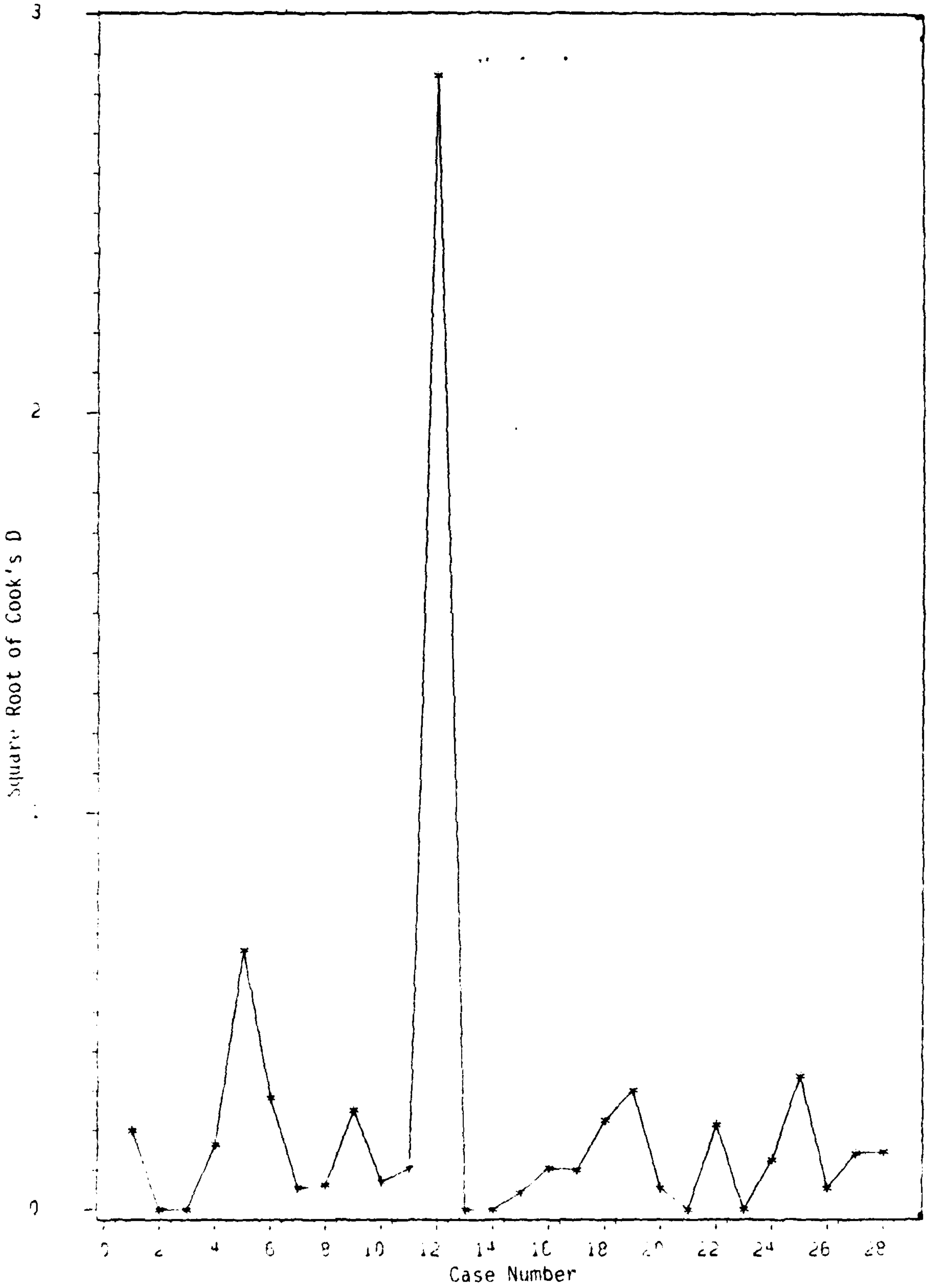
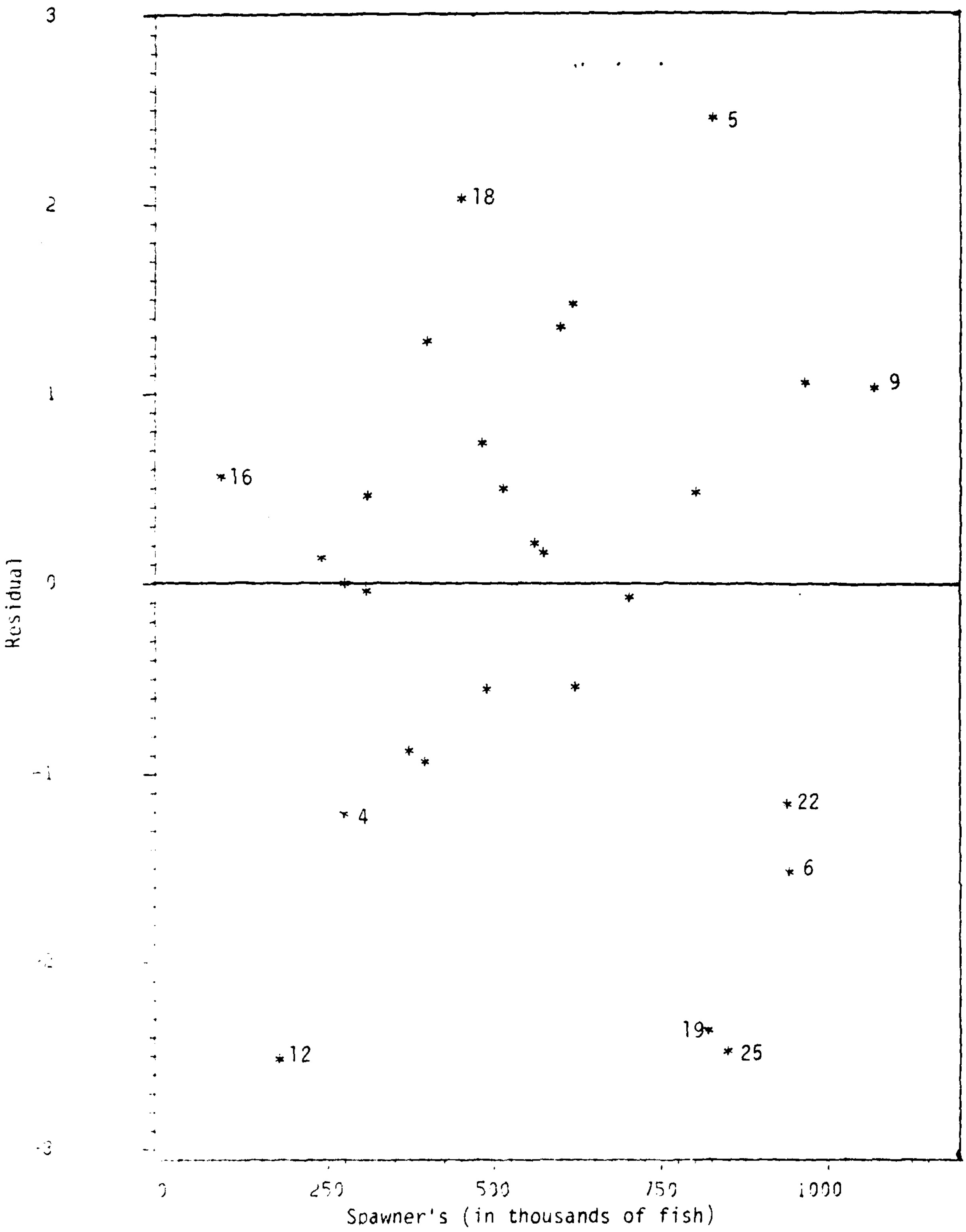

3600
3000
2400
1800
1200
600
0
Recruits (in thousands of fish)
0
400
800
1200
Spawners (in thousands of fish)
5
9
18
MEAN
MEDIAN
22
6
19
25
4
16
12

Figure 2

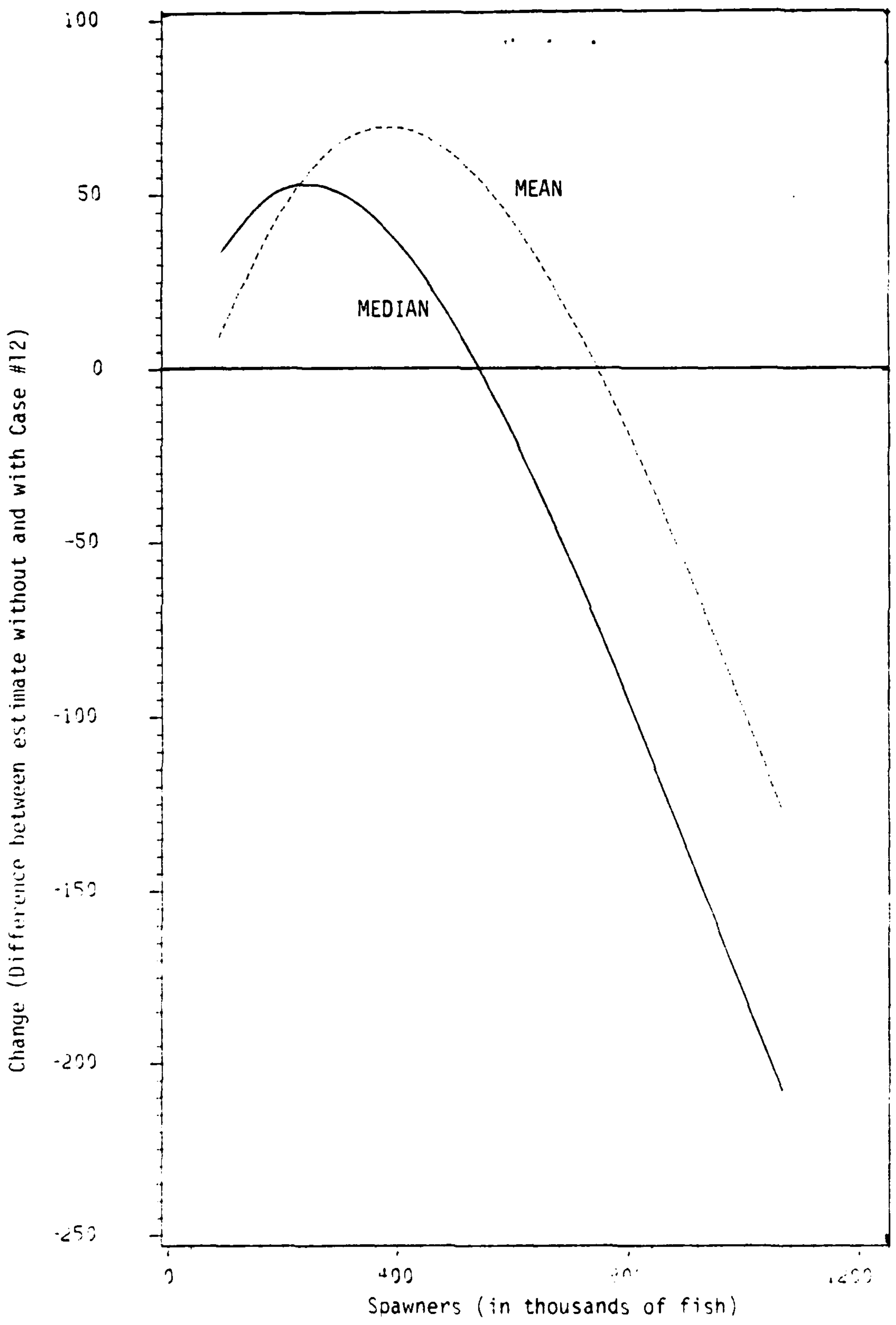

3
2
1
0
Square Root of Cook's D
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
Case Number

Figure



3
2
1
0
-1
-2
-3
Residual
5
18
9
16
22
4
6
19
25
12
0
250
500
750
1000
Spawner's (in thousands of fish)

Figure 4



100
50
0
-50
-100
-150
-200
-250
MEAN
MEDIAN
Change (Difference between estimate without and with Case #12)
0
400
800
1200
Spawners (in thousands of fish)

REFERENCES

ATKINSON, A. C. (1986). "Diagnostic tests for transformations." To appear in Technometrics.

BATES, D. M. and WATTS, D. G. (1980). "Relative curvature measures of nonlinearity." Journal of the Royal Statistical Society, series B, 42, 1-25.

BATES, D. M., WOLF, D. A., and WATTS, D. G. (1986). "Nonlinear least squares and first-order kinetics." Manuscript.

BELSLEY, D. A., KUH, E., and WELSCH, R. E. (1980). Regression Diagnostics. New York: Wiley.

BEVERTON, R. J. H., and S. J. HOLT. (1957). "On the dynamics of exploited fish populations." Her Majesty's Stationary Office, London. 533 p.

BOX, G. E. P. and COX, D. R. (1964). "An analysis of transformations (with discussion)." Journal of the Royal Statistical Society, Series B, 26, 211-246.

BOX, G. E. P. and HILL, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. Technometrics, 16, 385-389.

CARR, N. L. (1960). Kinetics of catalytic isomerization of n-pentane. Industrial and Engineering Chemistry, 52, 391-396.

CARROLL, R. J. (1982). Prediction and power transformation when the choice of power is restricted to a finite set. Journal of the American Statistical Association, 77, 908-915.

CARROLL, R. J. and RUPPERT, D. (1984). "Power transformations when fitting theoretical models to data." Journal of the American Statistical Association, 79, 321-328.

CARROLL, R. J. and RUPPERT, D. (1985). "Transformations in regression: a robust analysis." Technometrics, 27, 1-12.

COOK, R. D. and WANG, P. C. (1983). "Transformation and influential cases in regression." Technometrics, 25, 337-343.

COOK, R. D. and WEISBERG, S. (1982). Residuals and Influence in Regression. New York and London. Chapman and Hall.

DRAPER, N. and SMITH, H. (1981). Applied Regression Analysis, 2nd Edition. Wiley: New York.

DUAN, N. (1983). "Smearing estimate: a nonparametric retransformation method." *Journal of the American Statistical Association*, 78, 605-610.

HAMPEL, F. A. (1968). *Contributions to the Theory of Robust Estimation*. Ph.D. Thesis. University of California, Berkeley.

HAMPEL, F. R. (1974). "The influence curve and its role in robust estimation." *Journal of the American Statistical Association*, 62, 1179-1186.

HAMPEL, F. R. (1978). "Optimally bounding the gross-error-sensitivity and the influence of position in factor space." *1978 Proceedings of the ASA Statistical Computing Section*. ASA, Washington, D.C., 59-64.

HOAGLIN, D. C. and WELSCH, R. (1978). "The hat matrix in regression and ANOVA." *The American Statistician*, 32, 17-22.

KRASKER, W. S. (1980). "Estimation in linear regression models with disparate data points." *Econometrica*, 48, 1333-1346.

KRASKER, W. S. and WELSCH, R. E. (1982). "Efficient bounded-influence regression estimation." *Journal of the American Statistical Association*, 77, 595-604.

MILLER, D. M. (1984). "Reducing transformation bias in curve fitting. *The American Statistician*, 38, 124-126.

RAO, C. R. (1973). *Linear statistical inference and its applications, second edition*. Wiley: New York.

RICKER, W. E. (1954). "Stock and recruitment." *Journal of Fisheries Research Board of Canada*, 11, 559-623.

RICKER, W. E. and SMITH, H. D. (1975). "A revised interpretation of the history of the Skeena River sockeye salmon (Oncorhynchus nerka)." *Journal of the Fisheries Research Board of Canada*, 32, 1369-1381.

RUPPERT, D. (1985). "On the bounded-influence regression estimator of Krasker and Welsch." *Journal of the American Statistical Association*, 80, 205-208.

RUPPERT, D. and CARROLL, R. J. (1986). "Data transformations in regression analysis with applications to stock-recruitment relationships." To appear in the *Proceedings of the Ralph Yorque Conference of Natural Resource Management*. Springer: New York.

SNEE, R. D. (1985). "An alternative approach to fitting models when re-expression of the response is useful." Manuscript.

END

DTIC

10-86